

# THE LANCET

## Public Health

### **Supplementary appendix**

This appendix formed part of the original submission and has been peer reviewed.  
We post it as supplied by the authors.

Supplement to: Samuels EA, Taylor RA, Pendyal A, et al. Mapping emergency department asthma visits to identify poor-quality housing in New Haven, CT, USA: a retrospective cohort study. *Lancet Public Health* 2022; **7**: e694–704.

## Appendix Materials for ‘Mapping Emergency Department Asthma Visits to Identify Poor Quality Housing’

### Supplementary Methods

#### *Estimating population per parcel*

We first estimated the living area  $L$  ( $\text{m}^2$ ) =  $N \times A$ , where  $N$  is the number of stories, and  $A$  is the area of each residential parcel. Parcel size was computed using the ‘st\_area’ function from the R package ‘sf’. All parcels were assumed to be residential unless their use description was listed as otherwise in the tax assessor’s database (e.g., church, school, train station). The initial dataset included  $N$  only for parcels in New Haven city (52% of patients). By contacting tax assessor’s offices in North Haven, Guilford, East Haven, and Wallingford, we obtained information about the number of stories for those towns (69% of patients including New Haven). West Haven, Hamden, Branford, Madison did not have building stories data available. For these remaining parcels, we imputed missing values of  $N$  using a random forest classifier trained on the parcel value, area, structure style, use description, and zoning, which produced good fit to the training data ( $R=0.94$ , **Fig. S3A**). All parcels represented one residential building, except for four large HUD-subsidized complexes, one of which was Complex A, for which several component parcels were manually merged.

We then assigned population estimates for each parcel in the dataset by allocating the census block level population estimates equally among all residential parcels within a census block in proportion to their living area  $L$  (**Fig. S3C**), producing a distribution of estimates across all parcels (**Fig. S3B**). We confirmed the accuracy of our estimates using 62 HUD subsidized or owned parcels that had at least one ED visit for asthma over the study period. Number of visits and higher estimated occupants were strongly correlated ( $R=0.65$ ,  $p<0.001$ , **Fig. S3D**, **Fig. S3E**).

#### *Computing a composite measure of unique ED patients and total visits*

There were skewed distribution of visits per patient (**Fig. S1C**) caused by relatively rare high utilizers, examining total visits for a parcel can be misleading. Conversely, computing incidence rates using  $N_p$ , the number of unique patients (**Fig. S3G**, center) ignores repeat ED visits from the same patient, which may be caused by recurring indoor exposures from poor quality housing. We therefore used a linear mixed model to estimate the number of visits per patient ( $N_v$ ) for each parcel, which are shrunken towards the mean, diminishing the effect of rare high utilizers. We then calculated a composite asthma ED utilization incidence rate, which captures both number of patients and numbers of visits, defined as the ratio of the product  $N_v \times N_p$ , to  $\log_2(P)$ , where  $P$  is the estimated population per parcel. This incidence rate, used as the crude incidence rate for the remainder of the study, showed an even stronger relationship with housing conditions assessed by REAC scores (**Fig. S3G**, right).

#### *Sensitivity analysis of number of patients per housing unit*

The fewer patients observed visiting the ED from a given building, the less information the model can use to estimate the effect of that building on the probability of an ED visit for asthma and may reduce model performance. We therefore re-ran the model excluding buildings where fewer than  $n$  visits were observed, where  $n$  ranged from 1 to 10 and observed increasing agreement between asthma incidence rates and REAC inspection scores (**Fig. S6**).

#### *Statistical Analysis*

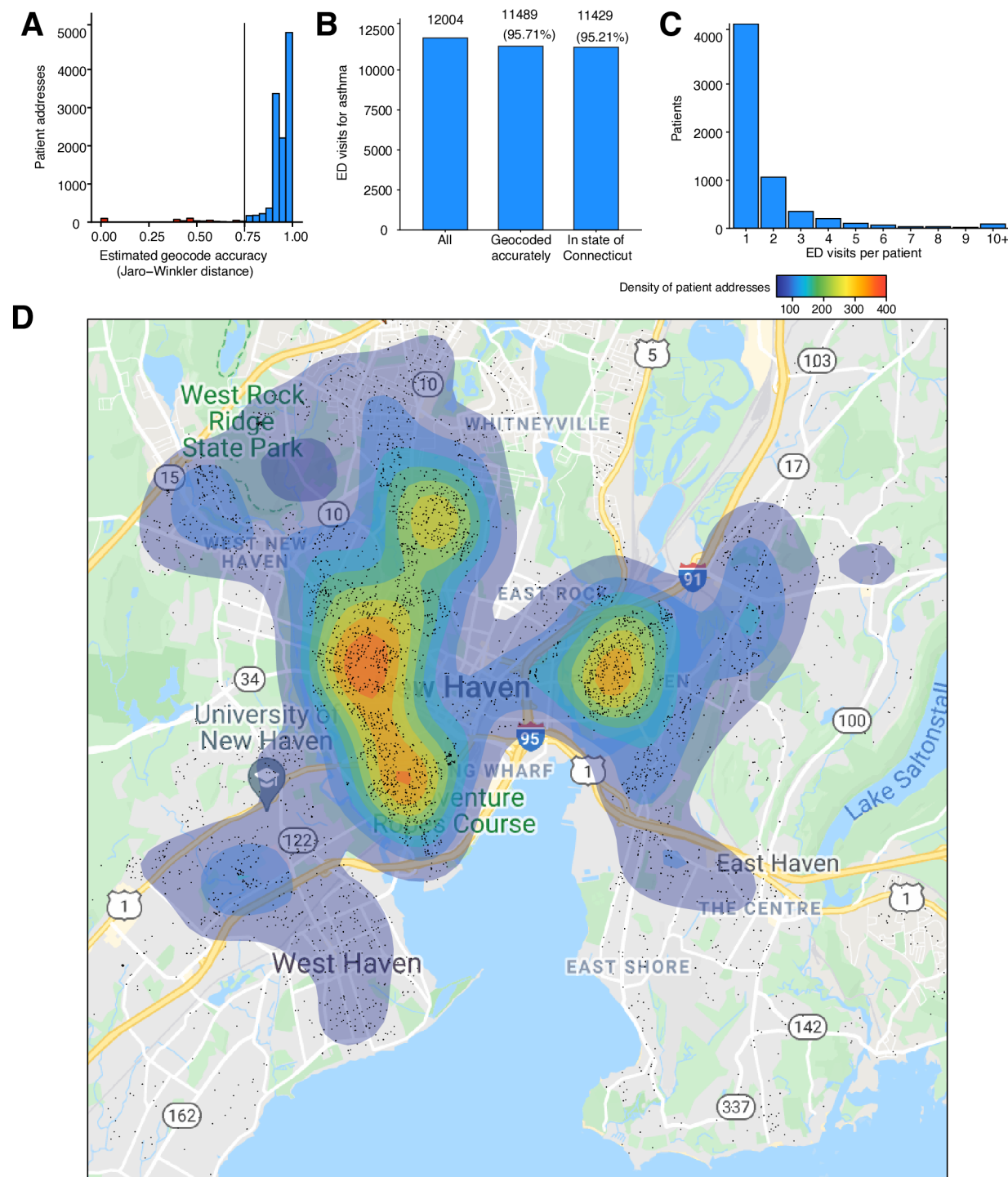
All statistical analysis were performed using R version 4.0.3 and finalized by September 2021. We fit linear and random forest regression models to estimate residual asthma burden after adjusting for the mix of people living in each housing complex. We fit random forest models using the ‘ranger’ package and linear

mixed models using ‘lme4’. Geocoding was performed using the package ‘ggmap’. To improve accuracy of estimating number of stories per parcel, missing values of zoning, structure type, and estimated parcel value were imputed using  $k$ -nearest neighbor imputation using the ‘VIM’ package<sup>1</sup>.

### **Additional References**

[1] Kowarik A, Templ M (2016). “Imputation with the R Package VIM.” *Journal of Statistical Software*, **74**(7), 1–16. doi: [10.18637/jss.v074.i07](https://doi.org/10.18637/jss.v074.i07).

Figure S1



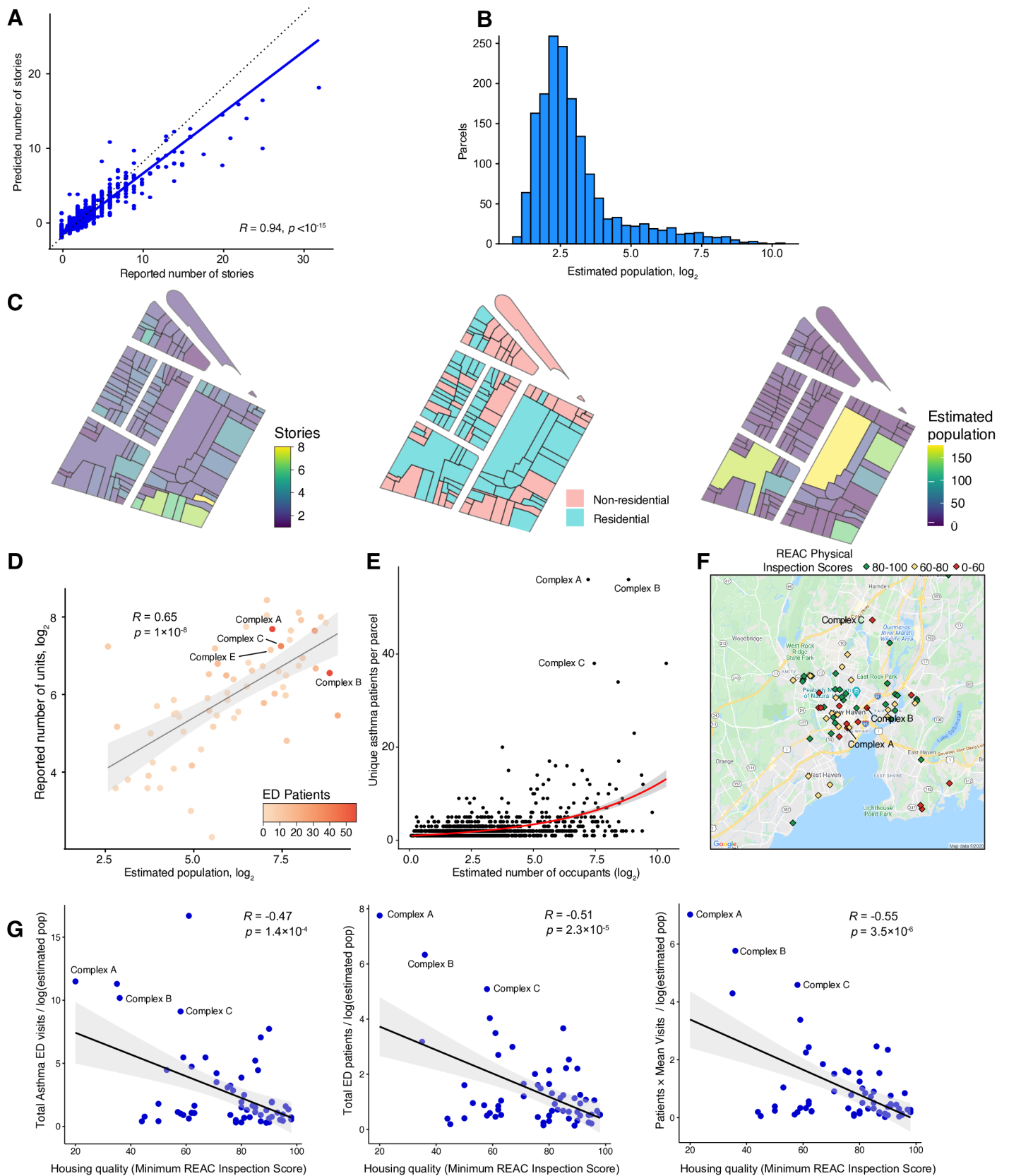
**Supplemental Figure 1.** Mapping emergency department visits for asthma using geospatial analysis. A. Bar plot shows a histogram (y axis) of geocoding accuracy (x axis) estimated by comparing the patient address to the inferred address returned by the Google Maps Geocoding API using the Jaro-Winkler string similarity metric. Only geocodes with very high similarity (Jaro-Winkler similarity > 0.75, 95.7%) are retained for further analysis. B. Bar plot shows the total number of ED visits (y axis) before and after geocoding, as well as after removing out of state addresses. C. Histogram of number of ED visits for each individual patient. D. Geospatial distribution of patient addresses based on geocoded coordinates (black points), overlaid on a street map of central New Haven. Contours show a kernel density estimate (color bar).

# Figure S2



**Supplemental Figure 2.** Assigning patient addresses to multi-address parcels. **A.** Schematic drawing (adapted from<sup>34</sup>) of the layout of multi-building subsidized housing Complex A. Each of the different buildings in the complex (color legend) has a different address. **B.** Rendering of the parcels near Complex A from the New Haven parcel dataset, overlaid with the geocoded locations of patient addresses (points), which are linked with the closest parcel using a nearest-neighbor join.

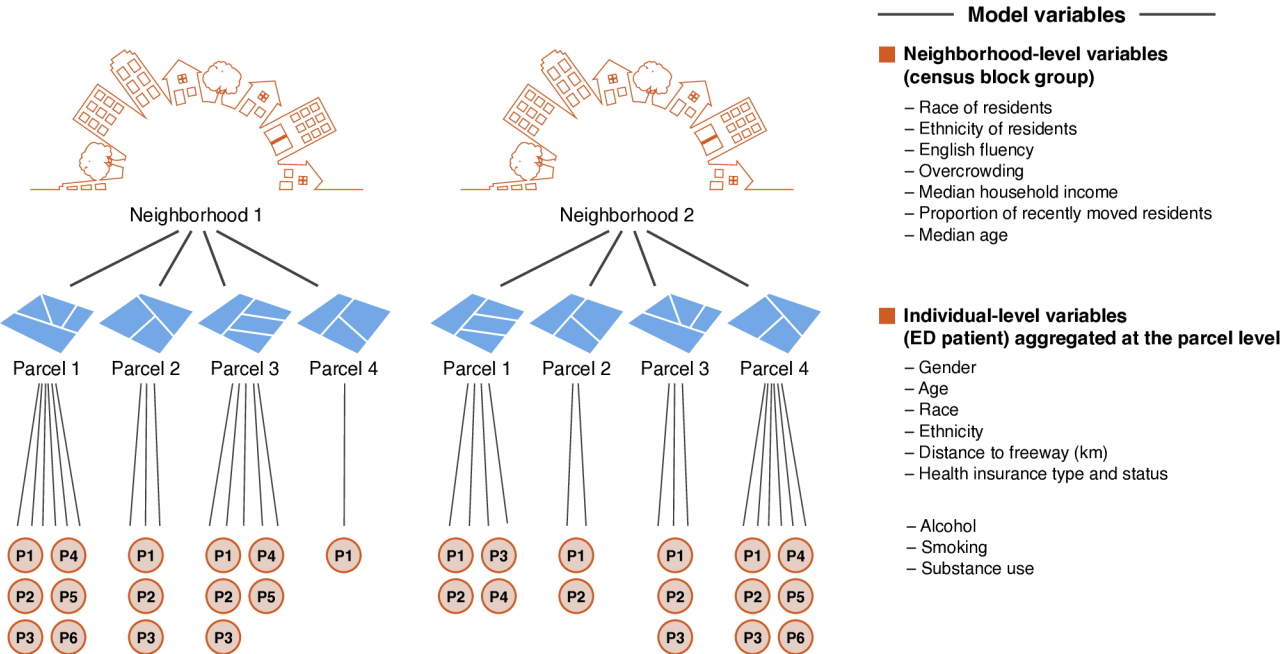
**Figure S3**



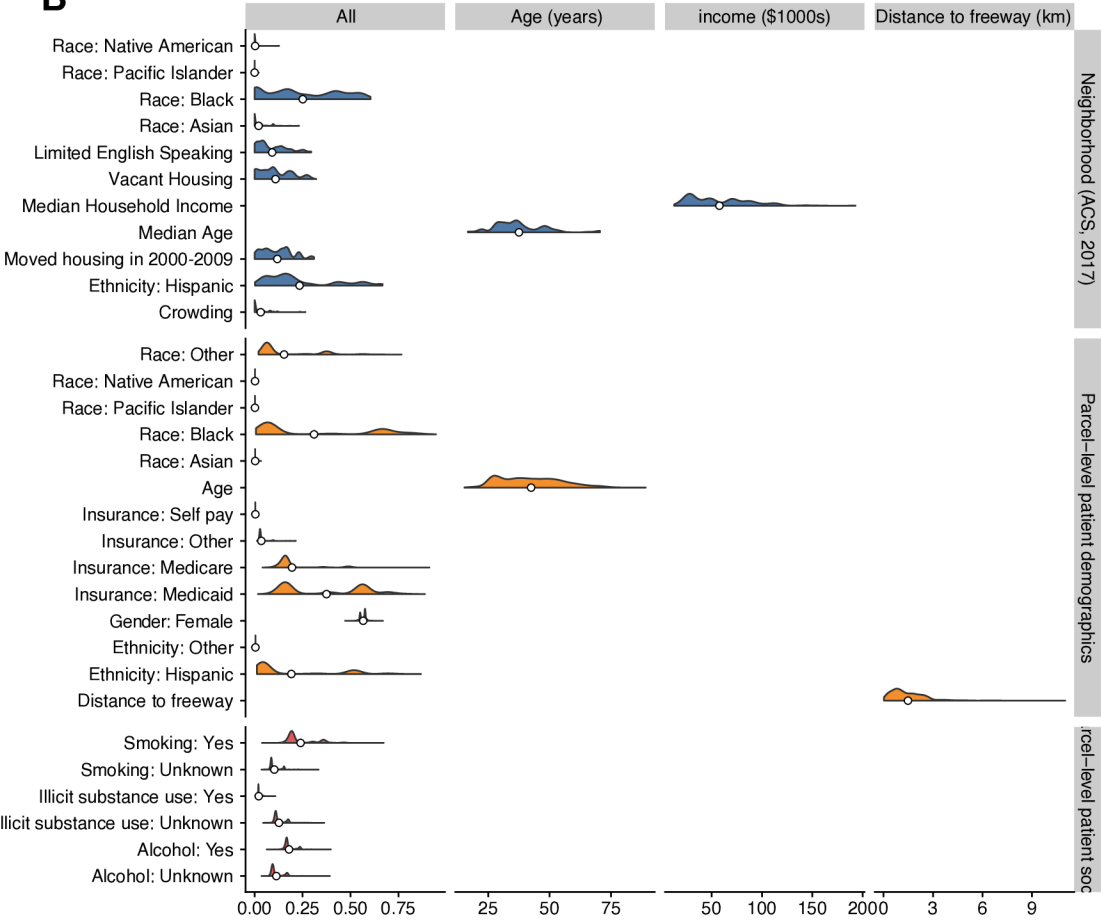
**Supplemental Figure 3. Estimating asthma ED utilization at the parcel level.** **A.** Scatter plot shows the predictive performance (training set error) of a random forest classifier designed to predict the number of stories (y axis) for each parcel based on parcel characteristics (value, zoning, structure style). **B.** Histogram of estimated population using inferred number of stories and parcel area. **C.** Visual schematic of the algorithm to estimate the parcel-level population distribution shown in B. The livable area for each census block (total area X number of stories) is computed and the population is distributed equally to each residential parcel in the block group based on the proportion of living area it makes up. Schematic is a rendering of all parcels in a randomly selected block group in central New Haven. **D.** Validation of population estimates (x axis) based on reported numbers of units in HUD-owned and subsidized housing in New Haven (y axis). Color legend shows the number of patients tracks with both estimated and reported size, providing a second layer of validation. **E.** Scatter plot shows the relationship between estimated population per parcel (x axis) and observed asthma ED patients (y axis). Poisson model fit also shown (red line). **F.** Geospatial distribution of HUD-owned and subsidized housing in New Haven. Each housing complex is shown on the map by a point marking its location after geocoding and a color indicating the outcome of its most recent REAC inspection (legend). **G.** Alternative asthma ED utilization rate metrics (y axis) compared to REAC inspection scores (x axis) for 62 HUD-owned and subsidized complexes in New Haven (points).

Figure S4

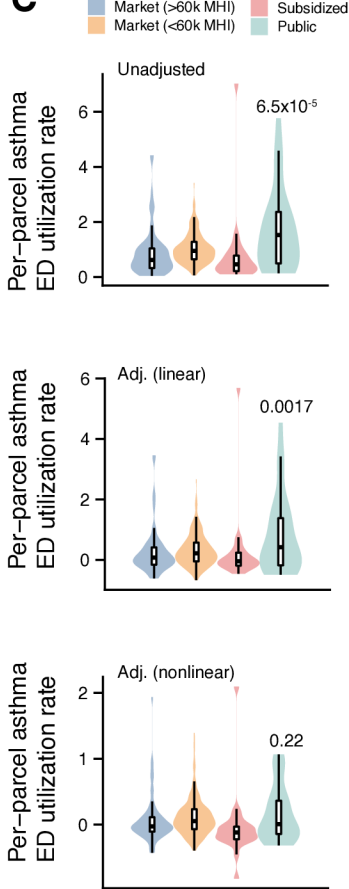
A



B

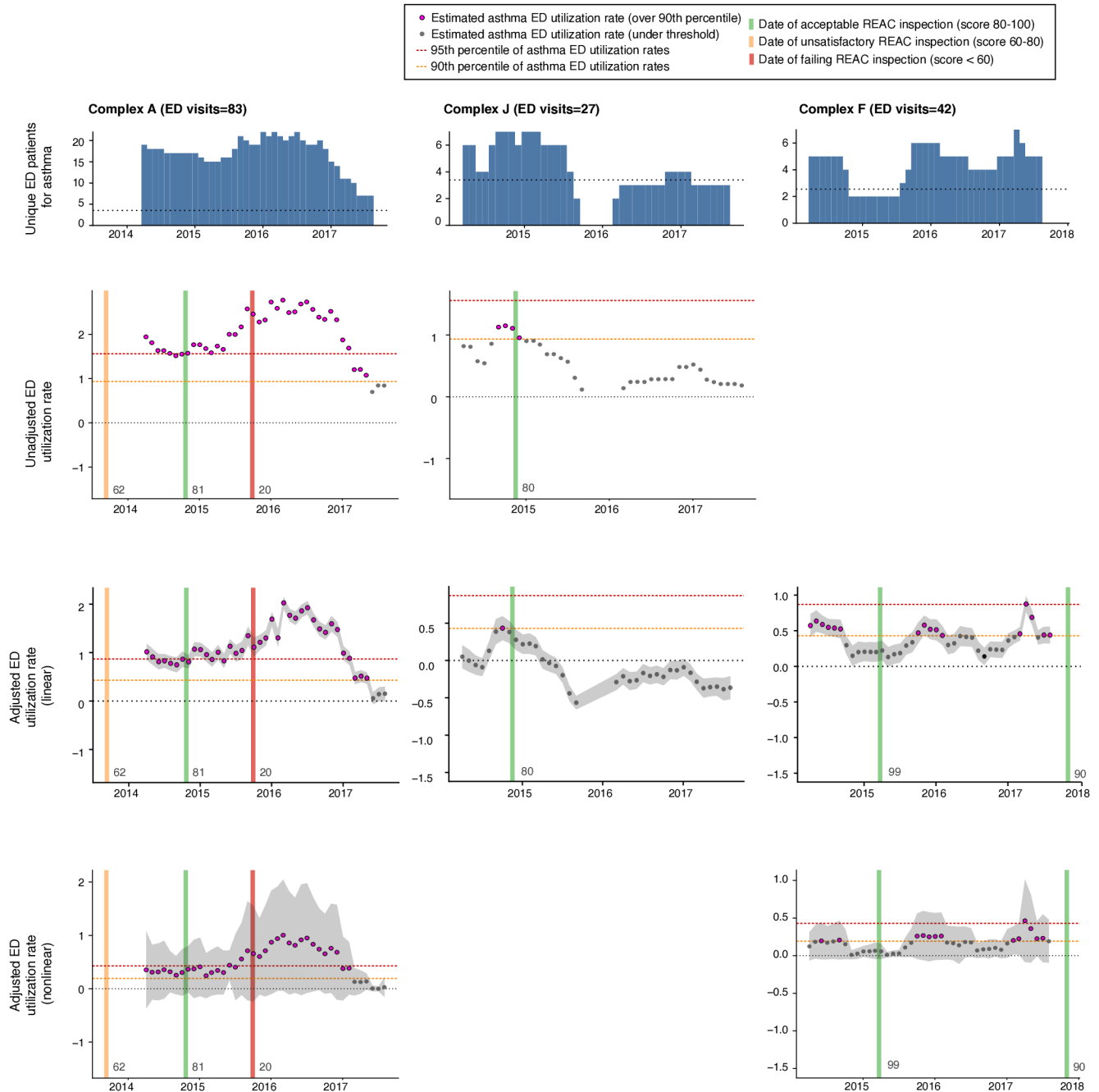


C



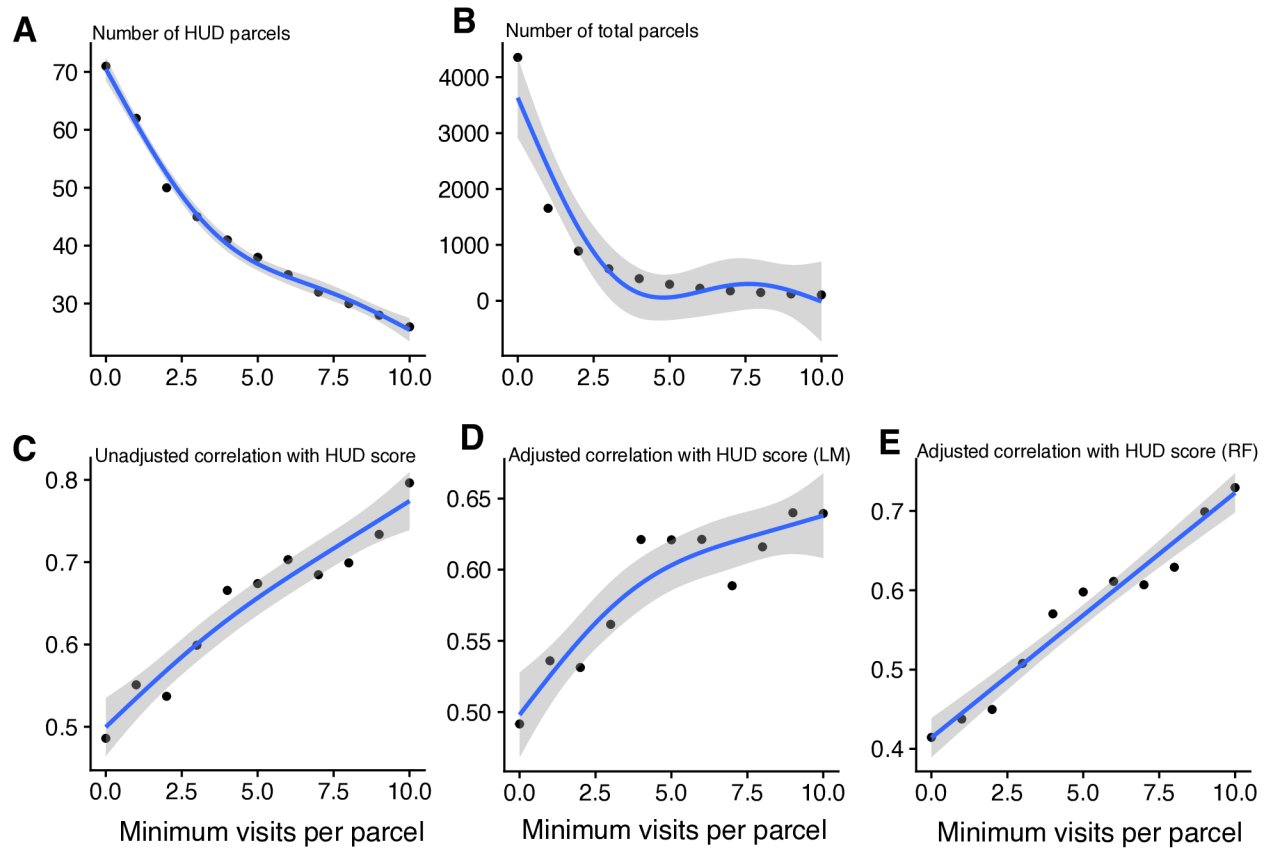
**Supplemental Figure 4. Controlling for individual and neighborhood-level demographics using regression models.** **A.** Schematic of regression model structure and variables. The model incorporates two types of predictor variables: neighborhood, and individual patients aggregated at the parcel level (left panel). All predictors are passed to the regressions in an equivalent manner, regressions are not hierarchical. **B.** Distributions of all neighborhood variables used in the analysis. Values were obtained at the census block group level from the American Community Survey (ACS), 2017. Age and median income values were scaled to the range 0-1. **C.** Differences in asthma ED visit incidence rates by type of housing (color legend). Market-rate housing is split by neighborhoods with average annual MHI above and below \$60,000. Box plots show the median and interquartile range, *p*-value: Wald test, MHI: median household-income.

# Figure S5



**Supplemental Figure 5.** Temporal distribution of asthma ED utilization for the three complexes classified as false positives. Complex J and F (right columns) were only classified as false positives for two of the models, so only two of the three rows are populated. See legend for Figure 3 for further details.

**Figure S6**



**Supplemental Figure 6.** Sensitivity analysis for the effect of filtering by the minimum number (n, x-axis) of observed ED visits per parcel. **A.** The number of HUD parcels available with inspection scores and at least n visits. **B.** The total number of parcels with at least n ED visits. **C-D.** Model performance, before (C) and after (D) as determined by correlation (y-axis) of asthma ED utilization incidence rates with HUD scores.

## Appendix Table 1. Adjusting estimates of asthma ED utilization incidence.

Coefficient estimates are given first, with 95% confidence intervals in parentheses, followed by *p* values. Predictors refer to individual patients aggregated at the parcel level, unless neighborhood is specified, in which case predictors are from American Community Survey at the block group level.

Distance to Freeway (km)	−0.027 <sup>+</sup> (−0.056, 0.001) p = 0.063
Sex: Female	2.565*** (1.457, 3.674) p = 0.00001
Race: Black or African American	0.307*** (0.208, 0.406) p = 0.000
Race: Other	0.259* (0.026, 0.493) p = 0.030
Race: Asian	−12.368 (−27.942, 3.205) p = 0.120
Ethnicity: Hispanic	0.270** (0.104, 0.436) p = 0.002
Ethnicity: Other	−44.333 (−126.487, 37.821) p = 0.291
Insurance: Medicare	0.053 (−0.183, 0.289) p = 0.661
Insurance: Medicaid	0.360*** (0.233, 0.488) p = 0.00000
Insurance: Other	1.733*** (0.740, 2.725) p = 0.001
Insurance: Selfpay	10.369 (−16.325, 37.063) p = 0.447
Smoking: Yes	0.494*** (0.243, 0.744) p = 0.0002
Smoking: Unknown	0.462 (−0.688, 1.611) p = 0.431
Alcohol: Yes	−0.451 (−1.121, 0.220) p = 0.188
Alcohol: Unknown	1.799 <sup>+</sup> (−0.047, 3.645) p = 0.057
Illicit: Yes	1.708 (−0.711, 4.127) p = 0.167
Illicit: Unknown	−0.306 (−2.240, 1.629) p = 0.757
Mean Age (Years)	−0.004** (−0.006, −0.001) p = 0.002
Neighborhood Race: Black or African American	−0.042 (−0.224, 0.141) p = 0.655
Neighborhood Race: Asian	0.073 (−0.760, 0.907) p = 0.864
Neighborhood Ethnicity: Hispanic	−0.289** (−0.483, −0.096) p = 0.004
Neighborhood Median: Age	−0.003* (−0.006, −0.001) p = 0.021
Neighborhood Median Household Income	0.001* (0.00002, 0.002) p = 0.047
Neighborhood Crowding	−0.105 (−0.573, 0.364) p = 0.661
Neighborhood Vacant Housing	0.237 (−0.103, 0.577) p = 0.173
Neighborhood Limited English Speaking	0.001 (−0.450, 0.451) p = 0.998
Neighborhood Moved Housing 2000-2009	−0.297 <sup>+</sup> (−0.632, 0.037) p = 0.082
N	1654
R-squared	0.189
Adj. R-squared	0.175
Residual Std. Error	0.438 (df = 1625)
F Statistic	13.539*** (df = 28; 1625)
+p<0.1; *p<0.05; **p<0.01; ***p<0.001	